# 聊天机器人发展历程及构建方法-简介

**摘要**  本文中梳理总结近些年关于聊天机器人的发展历程，分析上个世纪第一个聊天机器人系统问世到当今最为流行的实现技术。同时对聊天机器人的用途、分类、系统框架、构建方法、评测标准以及存在问题进行了介绍。最后通过对现有技术存在的问题进行分析，展望未来的发展方向。

**关键词 聊天机器人；seq2seq；生成式；检索式**

**引言**

聊天机器人是目前非常热的一个人工智能方向。聊天机器人通过学习自然语言来与人类进行对话，属于问答系统的范畴[1]。21世纪互联网的爆发对聊天机器人创造了更大的发展空间，研究者们从互联网上获取了海量的聊天数据，同时计算机硬件能力的提升也为新一代的深度学习算法提供了更好的平台。目前，深度学习技术在自然语言处理NLP、图像识别等领域中已经取得了显著的成就，在自然语言处理方面如信息抽取、自动摘要、语义分析、词性标注已经超过了传统方法，也就说明构建聊天机器人时可以通过深度学习技术对上下文内容进行分析生成一个满意的回复[2]。

聊天机器人潜在的巨大商业价值和社会价值使之日益成为工业界和学术界的研究热点[3]。

## 1 国内外发展历程

图灵于 1936 年提出图灵机（Turing Machine）这一模型，为今后的计算机制造提供了理论基础，因而也被冠以“人工智能之父”的称号。1950 年他发表《计算机器与智能》一文，并在文中提出了机器具有智能的定义，这一标准被人们称为图灵测试[4]。图灵测试给聊天机器人种下了一颗种子。从上世纪60s到上世纪末，聊天机器人大约经历了三个重要的历史时期。

      第一个时期，1966年麻省理工学院（MIT）的约瑟夫·魏泽鲍姆（Joseph Weizenbaum）开发的聊天机器人ELIZA[5] 为世界第一款聊天机器人，它的主要用途是辅助医疗，用于在临床治疗中模仿心理医生。ELIZA的实现技术为关键词匹配及人工编写的回复规则，它并不理解对话内容且知识范围有限，只能在特定领域内使用。

      第二个时期，1988年加州大学伯克利分校（UC Berkeley）的罗伯特·威林斯基（Robert Wilensky）等人开发了名为UC（UNIX Consultant）[6]的聊天机器人系统。UC是一款帮助用户学习怎样使用UNIX操作系统的聊天机器人。它具备了分析用户的语言、确定用户操作的目标、给出解决用户需求的规划、决定需要与用户沟通的内容、以英语生成最终的对话内容以及根据用户对UNIX系统的熟悉程度进行建模的功能。

      第三个时期，1995年受到ELIZA聊天机器人的启发，理查德·华勒斯（Richard S. Wallace）博士开发了ALICE系统。随着ALICE一同发布的人工智能标记语言AIML（Artificial Intelligence Markup Language）被广泛应用在商业客服的开发中。尽管ALICE采用的是启发式模板匹配的对话策略，但是它仍然被认为是同类型聊天机器人中性能最好的系统之一。

这些早期聊天机器人是使用音频或文本进行对话的计算机程序。这类程序的设计通常模拟人类作为聊天对象的行为，并以图灵测试作为是否成功的标准[7]。

到21世纪以来，聊天机器人随着互联网数据的爆发和计算硬件能力的提成以及新技术的背景下得到了飞速的发展。

在工业界，很多国内外顶级的互联网公司也纷纷推出了自己的产品，2010年苹果公司推出了Siri聊天机器人，通过理解用户的语音输入，可以帮助用户预定闹钟、播放音乐、查天气等信息，甚至帮助用户预订电影票。2012 年，京东自主开发完成了聊天机器人JIMI，能够向用户提供二十四小时的服务，服务内容涵盖电子商务的各个环节，如售前咨询、售后咨询等。2014年，微软公司在中国发布了Cortana和“小冰”，Cortana类似于Siri属于个人助理范畴。“小冰”注重聊天的情感联系，并把智商（IQ）和情商（EQ）考虑在内，以适当的人际关系反应来在长时间交互中吸引用户。通过对聊天过程中语境与语义的理解，“小冰”在一定程度上实现了超越简单人机问答的智能交互，能够使用更加类似人的交流方式[8]。2015年，阿里巴巴推出的聊天机器人产品阿里小蜜，它与 JIMI功能很相似，都是向用户提供购物环节中涉及的咨询类对话服务。同年，百度公司也推出了自己聊天机器人度秘。2017年小米公司也发布了自己的AI智能音响“小爱同学”，它支持语音交互可以播放音乐、电台点播、[小说](https://baike.baidu.com/item/%E5%B0%8F%E8%AF%B4/45851" \t "/Users/apple/Documents\\x/_blank)、脱口秀，教育学习、儿童类多种有声读物内容，它还可以控制小米电视、扫地机器人、空气净化器等一系列生态链设备。同年，阿里巴巴也推出了“天猫精灵”，他也小爱同学非常类似，可以提供感情陪护等一系列生活服务[3],丰富了人们的日常生活。

在学术界，早期AIML构建的检索式聊天机器人构建知识库时存在着大量人力资源和时间成本。对于构建特定领域的聊天机器人时需要使用大量的人工特征提取等工作，Serban[9]等人采用了基于循环神经网络的方法，省去了特征提取工作中的人工参与并取得了不错的效果。Google Vinyal[10]等人首次提出将序列到序列的模型用于对话系统，这是最早将机器翻译中的 sequence to sequence模型在对话系统中进行应用，该模型在机器翻译领域中将源语言句子进行编码，然后将目标语言句子进行解码，从而完成翻译任务，他将该模型应用于对话系统中，将问题句进行编码，回复句解码，从而进行对话生成任务。而后 Bahdanau[11]等人提出一种注意力模型 attention 机制，提高了翻译的准确性。Guo[12]等人引入深度强化学习的思想，将 DeepMind团队提出的DQN模型用于文本生成过程，他们的模型仍是沿用 encoder-decoder 框架，但是在 decoding 的过程中利用 Deep Q-Network 进行迭代探索生成结果。Li 等人[13]提出了一个强化学习框架，模拟两个 agent 让其自动对话，训练神经网络 seq2seq 模型，然后将 Encoder-Decoder 模型和强化学习整合，定义强化学习的 reward 保证对话的未来奖励最大，从而使对话轮数增加。在生成模型的训练过程中，除了传统的有监督模型训练，复旦大学的 Yu[14]等人提出了序列问题上的半监督学习算法，该论文提出了应用 GAN 对抗训练的方式结合强化学习算法进行文本生成，解决了传统 GAN 在 NLP 等序列问题上不能进行梯度更新的问题，为后面的 GAN 在 NLP 上的应用提供了思路。2017年，蒙特利尔学习算法研究所（MILA）发表了MILABOT[ ]聊天机器人研究成果，该聊天机器人采用强化学习方法，整合基于检索与生成的对话模型。相比国外，国内一些知名的高校以及科研机构对聊天机器人研究投入了很多，并在此领域获得了显著的成绩。比如，清华大学、中科院计算所、香港大学、哈尔滨工业大学等，其中，哈尔滨工业大学研发的benben[ ]提供高质量的对话服务。

目前对于蒙古语聊天机器人的研究还尚未起步，主要原因还是低资源语料。

## 2 聊天机器人分类

从功能上可以分为任务型和闲聊型聊天机器人，如苹果siri、微软Cortana、百度度秘等被称为个人助理，因为他们的任务是完成人们设定的指定任务，所以都属于任务性聊天机器人。而微软的小冰和小米公司的小爱等是属于闲聊型聊天机器人，因为此类机器人可以与用户无限制范围的交互。

从技术实现方方面又可以分为检索式和生成式。

检索式聊天机器人多数为任务型的特定领域服务。检索式聊天机器人需要事先准备对话数据库，聊天系统接收到用户输入句子后，通过在对话库中以搜索匹配的方式进行应答内容提取，很明显这种方式对对话库要求很高，需要对话库足够大，能够尽量多地匹配用户问句，否则会经常出现找不到合适回答内容的情形，因为在真实场景下用户的输入范围非常广泛，但是它的好处是回答质量高，因为对话库中的内容都是真实的对话数据，表达比较自然，不会出现无法错误，因此很适合用在特定领域的聊天机器人。基于检索式的聊天机器人虽然可以保证回复语句的流畅性，但是受限于语料本身，很多内容在构建的数据库中没有答案，此时聊天机器人便无法进行回答[8]。

生成式聊天机器人多数为闲聊型的开放领域服务。它采取技术思路与检索式不同，使用时下最为流行的深度学习技术从大量的对话语料中学出一套规则，在接收到用户输入句子后，自动生成回复内容，这种形式的机器人的好处是可能覆盖任意话题的用户问句，但是缺点是生成的回复质量很可能会存在问题，比如可能存在语句不通顺，存在句法错误等看上去比较低级的错误，无法保证每次的生成效果。

## 3 聊天机器人系统框架

实现人与计算机之间的自由交互越来越受到人们的重视。人们可以通过语音或者文本对聊天机器人进行指令控制。你可以对他下达指令，机器人则像个下属一样为我们提供服务。这大大节省了人们的时间成本，而且机器的记忆要比人类的记忆好很多，在机器人不确定问题如何回复的时候还可以利用搜索引擎查找相关答案提供给用户。这类机器人满足了人们对于未来生活的幻想，而且着实能够提高人们的生产效率[15曹东岩]。

一个完整的聊天机器人框架往往由语音识别模块、问答模块、语音合成模块组成 ，如图1所示。其中，问答模块是整个系统的核心部分，也是最难的部分。可以单独用来进行文本交互类的聊天机器人，对于系统智能程度的高低评估取决于问答理模块生成的答复。问答模块由自然语言理解、对话管理和自然语言理解组成。

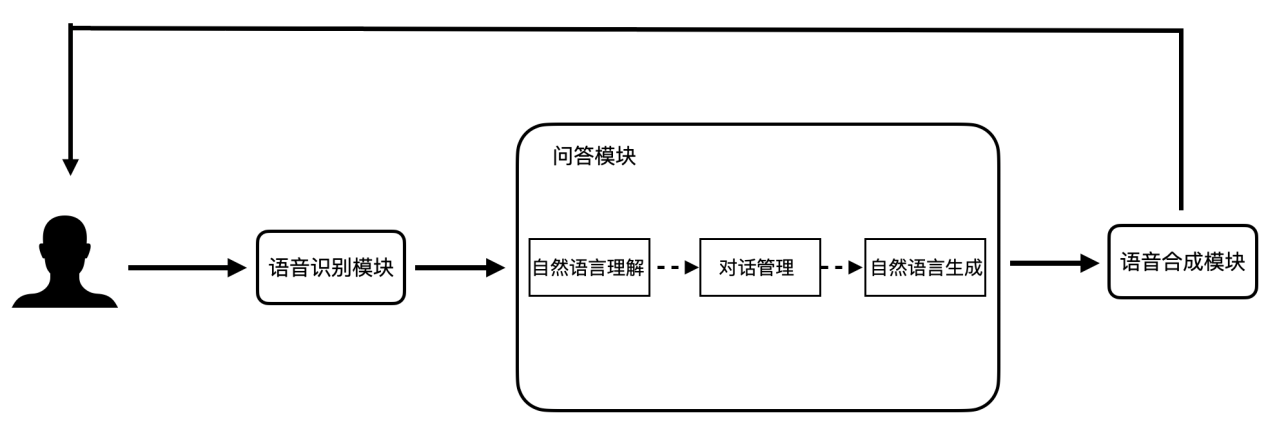


图 1 聊天机器人整体框架

聊天机器人的实现过程中首先通过语音识别模块将从用户获取的语音内容转换成文本内容，再把文本内容传送给问答模块。问答模块中的自然语言理解模块把文本内容转换成机器能够理解的数据，对话管理模块的功能是统筹自然语言理解模块、自然语言生成模块之间的协调工作，自然语言生成模块的功能就是根据用户提出的问题生成相应的回复，语音合成模块将从问答模块生成的最终文本内容合成为语音信展现给用户。

3.1语音识别模块

语音识别，通常称为自动语音识别，英文是Automatic Speech Recognition，缩写为ASR，主要是将人类语音中的词汇内容转换为计算机可读的输入，一般都是可以理解的文本内容，也有可能是二进制编码或者字符序列。但是，我们一般理解的语音识别其实都是狭义的语音转文字的过程，简称语音转文本识别（Speech To Text, STT）。

3.1.1语音识别技术发展历程

现代语音识别可以追溯到1952年，Davis等人研制了世界上第一个能识别10个英文数字发音的实验系统，从此正式开启了语音识别的进程。语音识别发展到今天已经有70多年，但从技术方向上可以大体分为三个阶段。

1993年到2009年，语音识别一直处于GMM-HMM时代，语音识别率提升缓慢，尤其是2000年到2009年语音识别率基本处于停滞状态；2009年随着深度学习技术，特别是DNN的兴起，语音识别框架变为DNN-HMM，语音识别进入了DNN时代，语音识别精准率得到了显著提升；2015年以后，由于“端到端”技术兴起，语音识别进入了百花齐放时代，语音界都在训练更深、更复杂的网络，同时利用端到端技术进一步大幅提升了语音识别的性能，直到2017年微软在Swichboard上达到词错误率5.1%，从而让语音识别的准确性首次超越了人类。目前在国内语音识别领域做的比较出色的由科大讯飞、百度等公司，其中科大讯飞的语音输入法在市场上应用广泛。

3.2问答模块

问答模块主要由两种技术来实现，首先对用户输入的问句进行数据处理，然后根据所属领域选择检索式还是生成式。其中检索式技术通过在对话数据库中以检索匹配的方式进行应答内容提取。生成式技术需要预先训练生成一个模型，再根据问句生成相应的应答句。

自然语言生成模块的功能就是生成用户回复。目前，实现聊天机器人对话系统的方法有两种：基于检索技术的对话系统以及基于生成技术的对话系统。基于检索式的对话系统是通过事先构建好的知识库，里面存放着问答对，用户输入一个问题，系统会根据句子相似度计算方法计算出知识库中相似度最高的问题，然后将这个问题的答案返回给用户，保证了对话的流畅性和准确性，不会出现语法错误和安全回答的现象，但比较依赖知识库，对知识库中完全没有的问题无法回答。基于生成式的对话系统则没有这方面的担忧，这种方式目前主要应用端到端框架，学习从问题到答案的关系，让机器人自己学会一套规则，能够自己逐字的生成回复。

3.3语音合成模块

最后使用语音合成技术将问答模块生成的应答句转换成语音信号反馈给用户。

3.3.1语音合成技术发展历史

语音合成技术大概经过了6个阶段的发展。

起源阶段，语音合成技术的起源可以追溯到18到19世纪，当时是用机械装置来模拟人的发声，那时候科学家们会制作出一些精巧的气囊和风箱去搭建发声的系统，可以合成出一些元音和单音。

电子合成器阶段，20世纪初，出现了用电子合成器来模拟人发声的技术，最具代表性的就是贝尔实验室的Dudley，他在1939年推出了名为“VODER”的电子发声器，使用电子器件来模拟声音的谐振。

共振峰合成器阶段，到了20世纪80年代，随着集成电路技术的发展，出现了比较复杂的组合型的电子发生器，比较代表性的KLATT在1980年发布的串/并联混合共振峰合成器。

单元挑选拼接合成阶段，到了20世纪80、90年代随着PSOLA方法的提出和计算机能力的发展，单元挑选和波形拼接技术逐渐走向成熟，90年代末刘庆峰博士提出听感量化思想，首次将中文语音合成技术做到了实用化地步。

基于HMM的参数合成阶段，在20世纪末期，还有另外一种基于HMM的参数合成技术出现。

基于深度学习的语音合成，随着AI技术不断发展，基于深度学习的语音合成技术逐渐被人们所知道，DNN/CNN/RNN等各种神经网络构型都可以用来做语音合成系统的训练，深度学习的算法可以更好地模拟人声变化规律。

## 4聊天机器人构建方法

4.1检索式聊天机器人构造

检索式机器人可以使用基于规则的方式进行模型的构造。

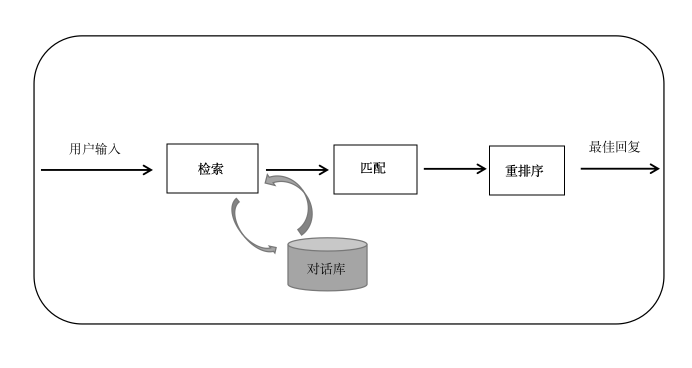


图3 流程框架如图3所示。

检索式聊天机器人中将预先准备的对话库，将用户输入的话语视为对该索引系统的查询，从中选择一个回复。具体来说，当用户在线输入话语后，系统首先检索并初步一批候选回复列表，再根据对话匹配模型对候选列表做重排序并输出最佳回复。

4.2生成式聊天机器人构造

计算机并不能直接理解人们的语言，需要将文本进行向量化。一般是做向量化处理也就是将文本内容上下文信息用向量来表示。将文本信息转化成向量格式有三种方法：第一种是 one－hot 编码，每一个词都被唯一编码成由 0 和 1 组成的向量，该向量仅有一个维度是 1，其余维度都是 0，向量长度为不同单词的总数量。这种表示方式有明显的缺点，一是向量维数非常大，二是无法刻画词与词之间的语义关系;第二种是文本分布式，其主要特点是维度比 one－hot 编码小，还可以表示一定的语义相关性，但是语义相关性比较浅层，且不易于扩展;另一种是文本向量空间模型(VSM)词嵌入的主流方式，这种方式将词映射到词向量空间，通过余弦函数可以得到两个单词之间的相关性，简单高效。

目前对基于生成式聊天机器人实现过程中主要采用Encoder-Decoder框架。所以本小节首先描述Encoder-Decoder框架技术原理。最早Vinyal等人将序列到序列的模型用于问答系统中。Encoder-Decoder框架可以看作是一种文本处理领域的研究模式，应用场景异常广泛，不仅仅可以用在对话机器人领域，还可以应用在机器翻译、文本摘要、句法分析等各种场合。图4是文本处理领域里常用的Encoder-Decoder框架最抽象的一种表示。

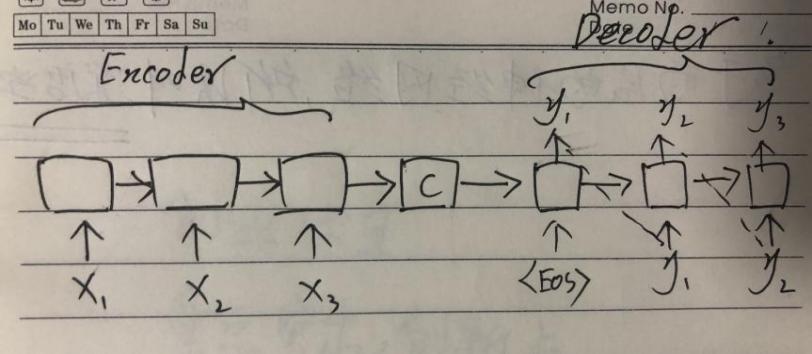


图4 2seq2seq 模型的基本结构

对于聊天机器人来说，完全可以使用上述的Encoder-Decoder框架来解决技术问题。具体而言，对应的<X,Y>中，X指的是用户输入语句，一般称作Message，而Y一般指的是聊天机器人的应答语句，一般称作Response。其含义是当用户输入Message后，经过Encoder-Decoder框架计算，首先由Encoder对Message进行语义编码，形成中间语义表示C，Decoder根据中间语义表示C生成了聊天机器人的应答Response。这样，用户反复输入不同的Message，聊天机器人每次都形成新的应答Response，形成了一个实际的对话系统。

神经网络在训练的时候容易产生过拟合，而且训练耗费大量的时间，为解决这个问题，Hiton等[17 ]提出 Drop out 机制。该机制定义为：在神经网络的训练过程中，按照一定几率将某些神经网络单元暂时从网络中丢弃，增强了模型的泛化能力。

在实际实现聊天系统的时候，一般Encoder和Decoder都采用RNN模型，RNN模型对于文本这种线性序列来说是最常用的深度学习模型，RNN的改进模型LSTM以及GRU模型也是经常使用的模型，对于句子比较长的情形，LSTM和GRU模型效果要明显优于RNN模型。尽管如此，当句子长度超过30以后，LSTM模型的效果会急剧下降，一般此时会引入Attention模型，2014年Bahdanau[11]等人提出了一种Attention机制，并取得了显著效果。具体框架如图5。我们注意到在原本的 seq2seq 模型中，解码器接收到的信息仅仅是编码器encoder 的最后一个输出向量 C，然后 RNN 网络随着时间的增加很可能原本的输入信息会丢失，这就造成了解码器最终效果不理想。Attention 机制能够很好的解决该问题。Google 研究人员在 Attention 机制基础上构建了Multi－Head的Attention 结构[18].

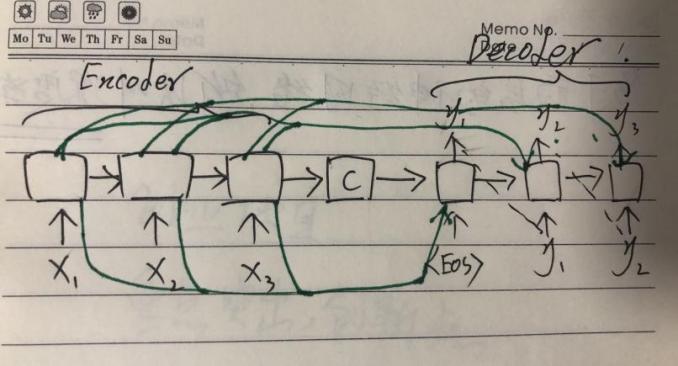


图 5 引入Attention机制后的seq2seq模型

百度的 Mou[19]等人提出基于关键词的预测，将关键词信息引入 decoder 部分，从而指导生成含有这些关键词的句子，解决了传统生成模型容易产生无信息的回复等问题。传统的 Seq2Seq 模型在解码过程采用的是 Gredy search，Gredy Search 本质上是一种贪心算法，它是在解码过程中每一个单词的输出都选择了概率最大的值，因为这样比较符合语言模型，但是有时候概率最大的值往往都是训练语料中出现概率最大的词，李纪为[20]等人把该类问题命名为“安全回复”问题。Beam Search算法则是在解码过程中输出概率最大的前K个单词，然后将这K个单词分别作为输入计算他们下一时刻的输出。最终能得到K个结果，所以系统的回复就能够变得更多样化。

## 4 聊天机器人模型评估方法及存在问题

通常情况下，用户评判一个智能聊天机器人或者问答系统的性能是否理想时，都会将该系统是否准 确地完成了用户的任务作为首要的参考条件，例如，在给定的对话中是否解决客户支持问题，但是在智能问答领域并没有一个合适的标准来衡量模型的性能，一般被经常使用的用于反映对话质量的方式主要 有如下几种。

4.1 BLEU

BLEU 最早用于评估机器翻译模型的质量[21]，也可以用来评估模型的生成序列和参考序列之间的 n元词组匹配度。目前被广泛应用在各种序列生成任务中，用于衡量 Seq2Seq 模型的效果。

4.2 句子相似度

句子相似度［22］是衡量文本之间相似程度的方式之一。句子相似度的计算可分两步进行，一是词向量加权平均获得句向量，二是根据句向量间的余弦相似度算法进行评估。

4.3 recall@k

    使用检索方式实现智能机器人，可以采用常用的检索型标准 recall@k，表示让这个模型从 10 个候选 响应中挑出 k 个最好的响应，候选的响应中包含 1 个真实响应和 9 个干扰项噪声响应。如果模型选择出 的 k 项回答中包含对应的正确响应，则该测试样本的结果将被标记为正确。自然，k 越大，那么这个任 务就会更加简单。

尽管我们知道一个对于回答多分类任务的语言模型的优化并不一定能作为一个好的语言回复生成模 型的好的参考标准，但是我们认为一个模型的分类能力的优化一定最终会带来生成任务的提升。

4.4 混淆度perplexity

      另一个常用于估测语言模型准确率的评估方法是 perplexity 困惑度，它被定义为每个单词的平均负对 数概率的指数[15]，如公式(1)所示。

      该指标可以反映“此模型对自己生成的目标序列的准确度是多少?”。更为精确的是，perplexity值 可以表达的一种概念是“如果我们在每个时间步从由该语言模型计算出来的概率分布中随机挑选单词， 要获得正确的答案，需要平均挑选几个单词?”。我们经常会将perplexity作为一种评估标准的参考选项， 是因为perplexity值越大，就表示模型之间的差距也更容易被人眼直观感知到。

4.3存在问题

最完美的机器人还是兼具任务型与闲聊型两者特点的：当我们需要个人助理的时候它能够为我们助力；当我们需要陪伴的时候，它有无限的时间与精力倾听我们的心声。---基于检索式和生成式模型的深度融合。

（1）目前聊天机器人仍然存在着多模态（文本、视频 、音频、图片）信息利用不足，情绪表达能力弱、模型融合机制缺失等问题。

（2）让聊天机器人拥有人类一样的思维，使得对话内容更加贴切。

（3）低语料条件下的聊天机器人构建。----少数民族语言

（4）目前对聊天机器人生成的对话内容评价方法还尚未成熟，有待提升。

（5）深入了解用户的目的、需求和历史信息。所谓的听懂人话。

**5展望**

蒙古语

**6 结语**

**参考文献**

[1]王浩畅,李斌.聊天机器人系统研究进展[J].计算机应用与软件,2018,35(12):1-6+89.

[2]胡艺钟. 检索与生成相结合的短文本对话研究[D].哈尔滨工业大学,2018.

[3]杨磊. 基于深度学习的聊天机器人研究与实现[D].北京邮电大学,2019.

[4] A. M. Turing. Computing Machinery and Intelligence. 1950, 59(236):433-460.

[5] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM, 1966, 9(1): 36-45.

[6] Wilensky R, Chin D N, Luria M, et al. The Berkeley UNIX consultant project[J]. Computational Linguistics, 1988, 14(4): 35-84.

[7]曹东岩. 基于强化学习的开放领域聊天机器人对话生成算法[D].哈尔滨工业大学,2017.

[8]Eliza 到小冰：社交对话聊天机器人机遇和挑战

[9]Serban I V，Sordoni A, Bengio Y,et al. Building end-to-end dialogue systems using generative hierarchical neural network model.[J] national conference on artificial intelligence,2016:3776-3783

[10]Vinyals O, Le Q. A neural conversational model[J]. arXiv preprint arXiv:1506.05869, 2015.

[11] BAHDANAU D，CHO K，BENGIO Y． Neural Machine Translation by Jointly Learning to Align and Translate［J］． Computer Science，2014( 3) : 1－15．

[12] Guo H. Generating text with deep reinforcement learning[J]. arXiv preprint

arXiv:1510.09202, 2015.

[13]Li J, Monroe W, Ritter A, et al. Deep reinforcement learning for dialogue

generation[J]. arXiv preprint arXiv:1606.01541, 2016.

[14]Yu L, Zhang W, Wang J, et al. SeqGAN: Sequence Generative Adversarial Nets

with Policy Gradient[C]//AAAI. 2017: 2852-2858.

[15 ]曹东岩. 基于强化学习的开放领域聊天机器人对话生成算法[D].哈尔滨工业大学,2017.

[16]孙立茹,余华云.基于深度学习的生成式聊天机器人算法综述[J].电脑知识与技术,2018,14(23):227-228.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[21] PAPINENI K，ROUKOSS，WARDT，et al．BLEU: A Method for Automatic Evaluation of Machine Translation［C］∥Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ( ACL) ，Philadelphia，2002: 311－318．

[22] PALAKORNA，HU X H，SHEN X J．The evaluation of sentence similarity measures［C］∥Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery，2008: 305－316．